Dynamic Optimized Advanced Scheduling of Bandwidth Demands for Large Scale Science Applications

Funded by US Dept. of Energy (DOE) Contract DE-FG02-09ER46622

Byrav Ramamurthy
Professor
Department of Computer Science and Engineering
University of Nebraska-Lincoln

Email: byrav@cse.unl.edu
http://cse.unl.edu/~byrav

DOE Annual PI meeting for the ASCR Network & Middleware
March 1-2, 2012

Outline

Our project involves scheduling network resources for large-scale scientific applications. We have made progress in the following areas:

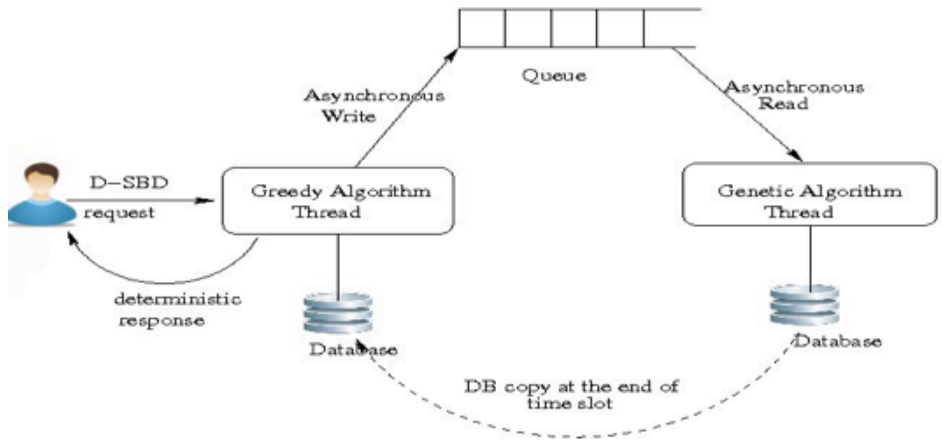
- **Re-optimization of dynamic bandwidth scheduling**: We designed a genetic algorithm to re-optimize the dynamic bandwidth scheduling to minimize blocking probability.
- **Network resource scheduling**: We designed an approach to jointly schedule computing, network and storage resources in Lambda Grid network with the goal of minimizing cost and total job completion.
- **DCN (Dynamic Circuit Network)**: We deployed DCN across a regional network spanning four universities (UNL, KU, KSU, UMKC) within GpENI (intra-domain) and between GpENI and MAX (inter-domain).
- **Green networking**: We proposed a local-optimized method to save energy in a core network comprising of composite/bundle links.
- **OpenFlow testbed**: We are deploying an OpenFlow testbed to connect the supercomputing resources and switches at the UNL campus.
- Related Projects: **NSF MobilityFirst Future Internet Architecture**.

Re-Optimization of Dynamic Bandwidth Scheduling

- Motivation
  - In practical network operation, many end users require deterministic services. Users make reservations in advance for end-to-end lightpaths for predefined service durations which called scheduled lightpath demand (SLD).
  - Dynamic-SLD cannot be precisely predicted and thus resource allocation for D-SLDs is difficult to optimize as a whole
  - User can reserve bandwidth at sub-wavelength level in dynamic scheduled bandwidth demand (D-SBD) to re-optimize dynamic bandwidth scheduling before a D-SLD is physically provisioned.

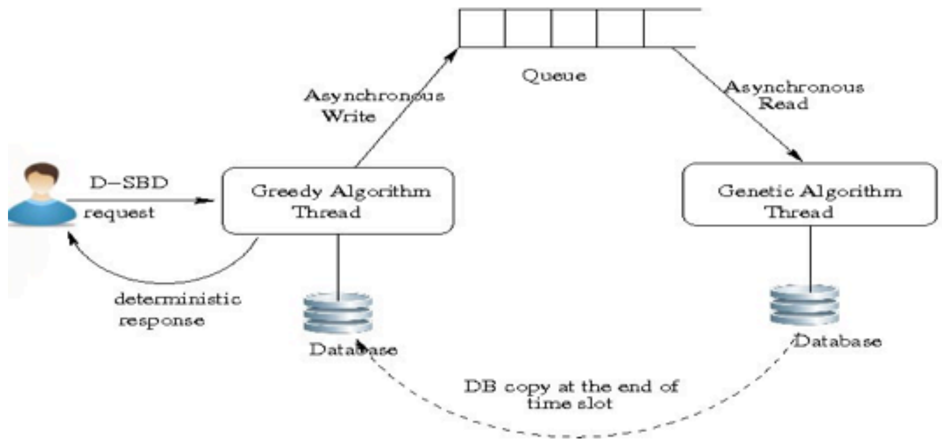
- Problem Statement
  - Given a D-SBD, provide a deterministic response to user quickly and also find best possible schedule to accommodate the maximum number of requests

Re-Optimization of Dynamic Bandwidth Scheduling (contd…)



- A continuous and parallel optimization method to address the dynamic and deterministic bandwidth scheduling problem in next generation wavelength-division multiplexing (WDM) networks.
- A greedy algorithm and genetic algorithm are run in parallel in separate threads and both of them take the Dynamic Scheduled Bandwidth Demand (D-SBD) as their input.
- The genetic algorithm takes as one of its inputs the output of the greedy algorithm and does the optimization of the D-SBDs with minimizing blocking probability as its fitness function. [1]

Re-Optimization of Dynamic Bandwidth Scheduling (contd…)

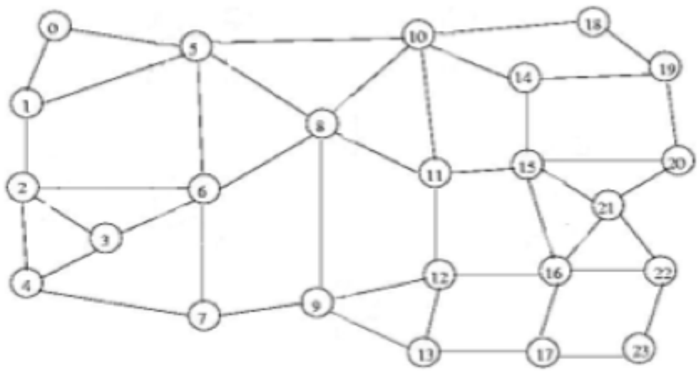


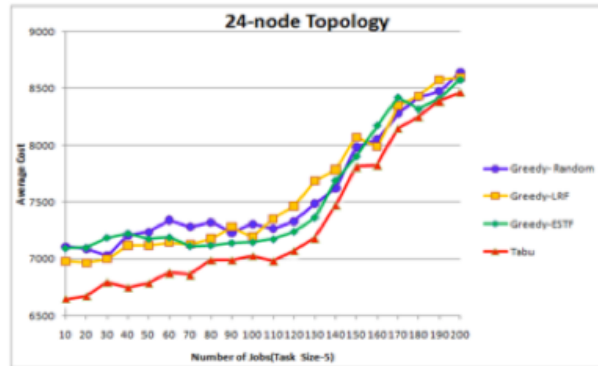
- Approach
  - Given a D-SBD, provide deterministic response to the user quickly and at the same time we want to find the best possible schedule
  - Greedy algorithm runs for a short amount of time and hence, can give a quick response.
  - Genetic algorithm which runs for a longer time finds the best possible schedule for a given set of D-SBDs.
  - Both the greedy algorithm and the genetic algorithm are run in parallel in separate threads.
  - Takes advantage of both to perform better as a system

Joint-Scheduling in Lambda Grid Network

- Motivation
    - Scheduling of the resources is one of the significant challenges in Grid computing and networking.
    - The objective of the joint scheduling is to schedule the resources effectively to maximize resource utilization, minimize cost, optimize completion time, etc.
- Problem statement
    - Joint optimization problems in an optical Grid network, which is jointly scheduling both computing resources and network resources.
    - Two objectives:
        - Minimizing the cost by proper use of the resource scheduling and routing
        - Minimizing the total completion time.

Joint-Scheduling in Lambda Grid Network





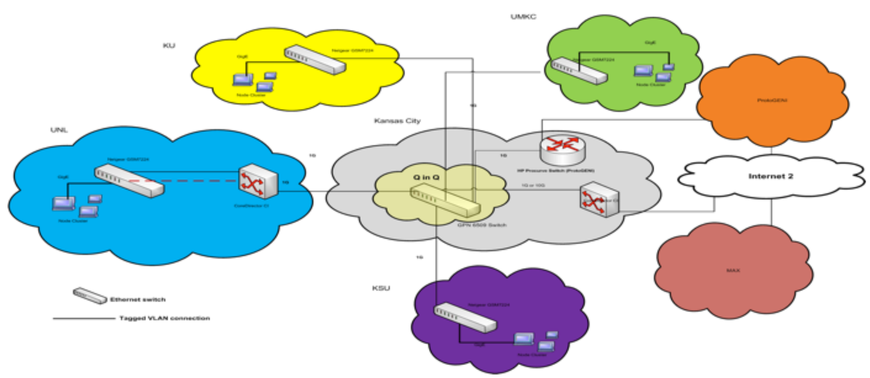
- Lambda Grid networks, are based on optical circuit switching and employ wavelength division multiplexing and optical lightpaths to provide a promising approach to create efficient infrastructure to support highly data-intensive demand high-performance computing applications.
- An approach based on Tabu Search heuristic for joint scheduling of computing, network and storage resources in a Lambda Grid network with the goal of minimizing cost and total completion time of job execution.
- Simulation results from both the methods show that the Tabu search heuristic performed better than the greedy approach in optimizing both the cost and completion time objectives. [2]

Joint-Scheduling in Lambda Grid Network (contd…)

- Conclusions
  - Tabu search approach performed at least 25% better than Greedy approach.
  - Tabu search- Avg cost behavior is better when job size is small in comparison with greedy.
  - Tabu search- Completion time is better when job size is large in comparison with greedy.
  - Greedy-ESTF performs better.
  - Tabu takes longer running times.

  - For small problem sizes, we can solve the optimization problem (using ILP solvers).

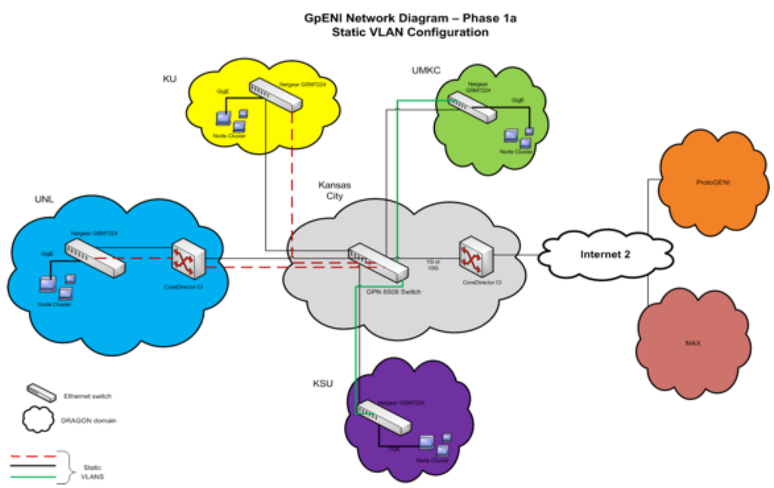
DCN Creation with PrairieFire & MAX



- Inter-domain DCN between GpENI and MAX
  - User can create circuit of fixed VLAN between Planetlab node in UNL and Planetlab node in MAX.
  - CMS data from the UNL Supercomputer (Prairiefire) was sent across to MAX over the dynamic circuit.
- Intra-domain across GpENI Universities
  - Users can create circuit of arbitrary VLAN (2-4094) between any of the 4 GpENI universities.
- In both cases, user can specify date, start time, end time, VLAN number using OSCARS web interface. [3][4]
- Funded by NSF/BBN GENI project “GpENI” 2008-2012.

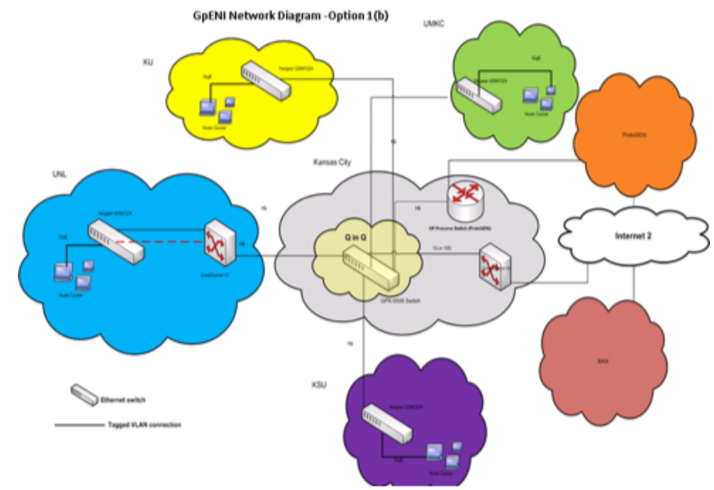
DCN Creation with PrairieFire & MAX

- Intra-domain DCN within GpENI



Static VLAN

- Configure static VLANs over the GPN switch so that DCN circuits can be created between universities only with the pre-configured VLAN tags.



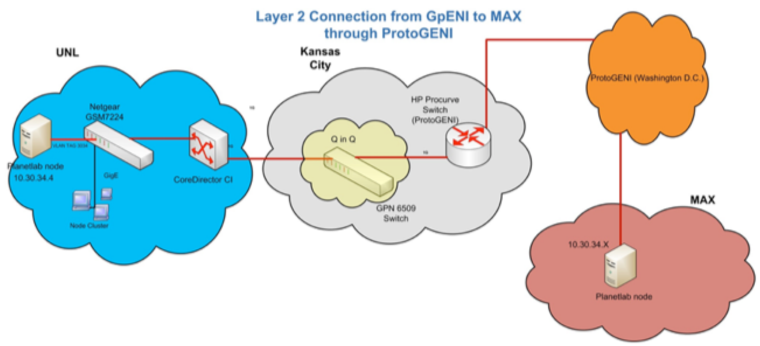
Q-in-Q

- Configure Q-in-Q cloud in the GPN switch with VLAN 125 so that it acts as a pass through for packets of any VLAN tag generated by any of the GpENI universities.

DCN Creation with Prairie Fire & MAX

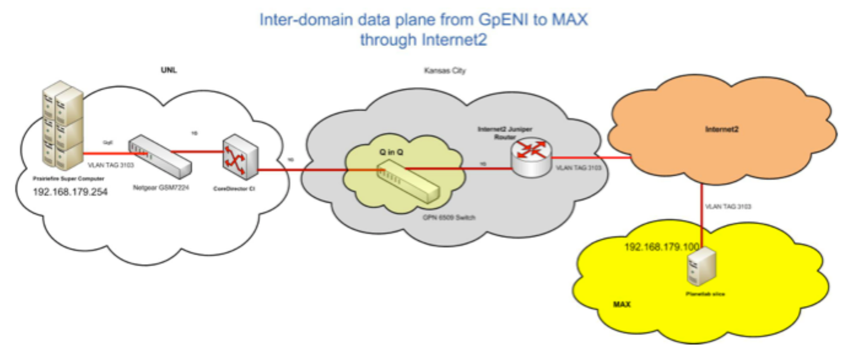
- Inter-domain DCN between GpENI and MAX

(Joint work by UNL Netgroup and Xi Yang & Tom Lehman (USC/ISI)



ProtoGENI

- Create static VLAN circuits (3031 to 3034) from HP switch to ProtoGENI facility at Washington D.C.
- Create DCN circuit (3034) from Washington D.C. to one MAX planetlab node. One planetlab machine at UNL was configured to receive packets of VLAN id 3034.



With CMS data (GpENI to MAX)

- Dynamic circuit established from UNL to MAX.
- Juniper port of Internet2 and UNL port connected in GPN switch using static VLAN.
- DCN established with this predefined VLAN.
- CMS Data from UNL transferred to MAX Planetlab machine.

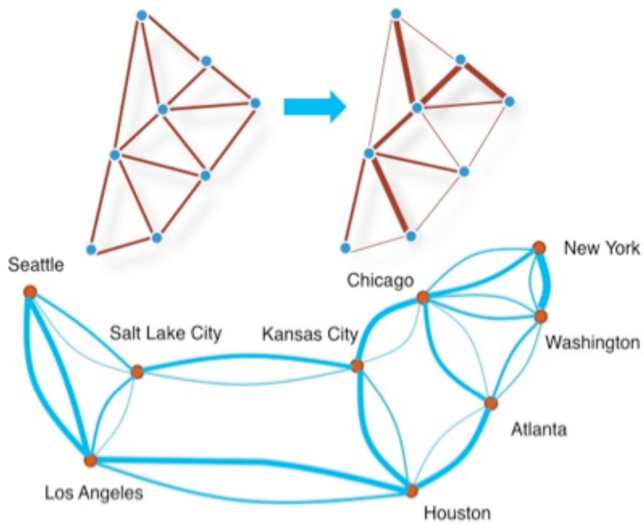
Byrav Ramamurthy, University of Nebraska-Lincoln

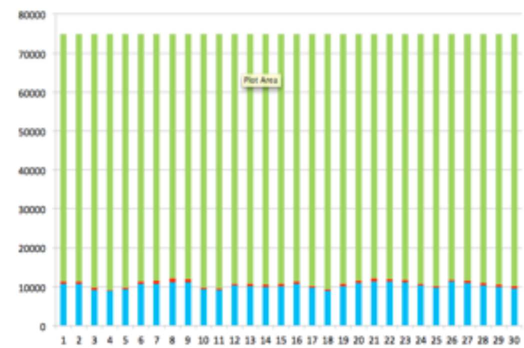
Saving Energy in the Core Network

- Motivation
  - Problems of globally-optimized green networking methods
    - Centralized, single point failure
    - ILP (Integer Linear Programming) problems, are not scalable
    - Slow acting
    - Require frequent changes to topology - undesirable
    - Exact traffic demand should be known in advance

- Objective
  - Design a local-optimized method which should be distributed, fast-acting, scalable, topology-stable.
- Key idea
  - Adjust bundle/composite link capacities by shutting down unused physical links and bringing them up during periods of high demand.

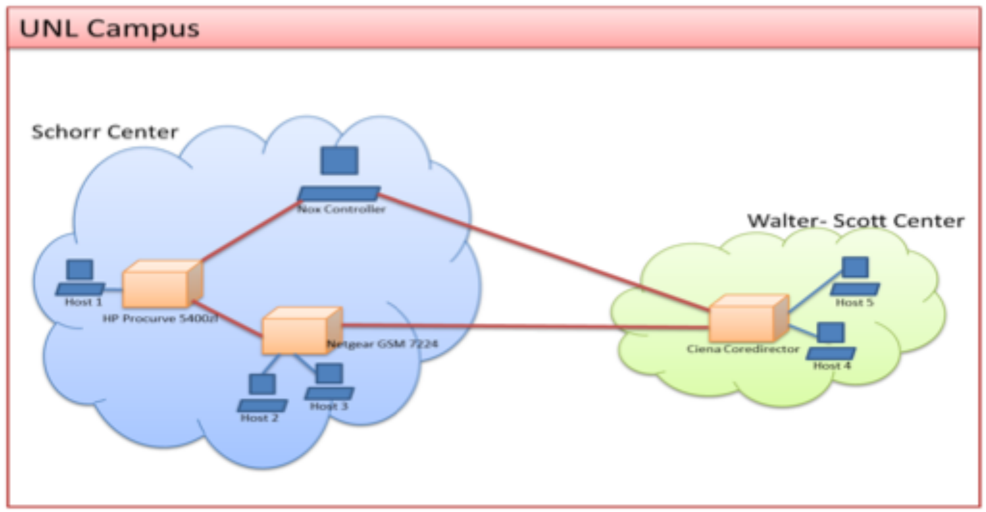
Saving Energy in the Core Network





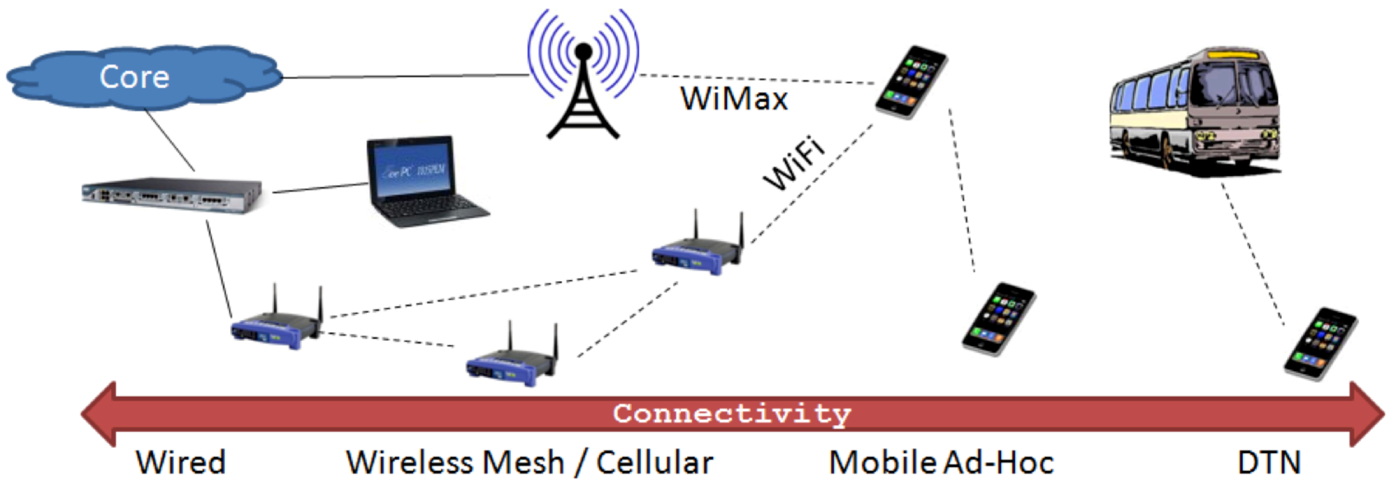
- A local-optimized method to save energy on ports of core routers by adjusting active sublinks within bundle link was developed. It has many advantages such as scalable, fast-acting and topology-stable.
- Internet2 based simulation experiment applying this method was conducted in both bin packing and load balancing cases.
- Results show > 80% port energy consumption of core routers are saved with limited increase of network congestion in the case. [5][6]

UNL OpenFlow Testbed



- UNL has two switches Ciena Coredirector and HP Procurve 5400zl running OpenFlow enabled firmware on them. They are intended to be controlled by an OpenFlow controller connected to them.
- A Netgear GSM 7224 is currently connected to the Ciena Coredirector.
- We are working towards establishing a working OpenFlow test bed at the UNL campus by connecting the switches together.
- Awaiting GENI rack deployment at UNL.
- We are also an NSF DYNES project participant.

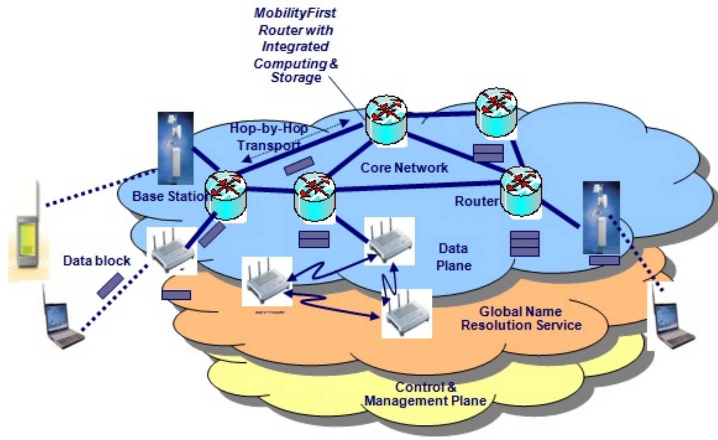
UNL MobilityFirst Project (funded by NSF FIA program)



Pic courtesy: D. Raychaudhuri, Rutgers

- Mobile devices are treated as the first class Internet citizens
- Shifting Intelligence into the network
- Decreases the emphasis on end to end setups as it has been with connection-oriented TCP/IP protocols.

UNL MobilityFirst Project



Pic courtesy: D. Raychaudhuri, Rutgers

Considering ‘mobility’ as a norm, the future internet is expecting an upsurge of traffic and a replacement of the fixed-host/server model. MobilityFirst architecture is designed to meet these challenges. The major design goals of the architecture are:

- Mobility, with dynamic host and network mobility at scale
- Robustness with respect to intrinsic properties of the wireless medium
- Trustworthiness in the form of enhanced security and privacy for both mobile networks and wired infrastructure
- Usability features such as support for context-aware pervasive mobile services, evolvable network services, manageability and economic viability
- *Need for fast optical paths between MobilityFirst GDTN (Generalized Delay Tolerant Networking Routers)*

Conclusions

Our project is developing theoretical and experimental approaches towards enabling a hybrid core network which can support dynamic bandwidth scheduling.

- Re-optimization of dynamic bandwidth scheduling
- Network resource scheduling
- DCN (Dynamic Circuit Network)
- Green networking
- OpenFlow testbed
- NSF MobilityFirst Future Internet Architecture.

Future Challenges

- Implementing re-optimization algorithms using OSCARS 0.6/ ARCHSTONE framework
- Extending the capability of Dynamic Circuit Networks into regional networks and campuses ("end-to-end") with science experiments
- Applications to carrier's core networks (beyond R&E & Govt. and into commercial networks, such as AT&T) (**ongoing collaboration with AT&T Labs-Research** [7])

References

[1] Pragatheeswaran Angu and Byrav Ramamurthy, "*Continuous and Parallel Optimization of Dynamic Bandwidth Scheduling for WDM Networks*," IEEE **GLOBECOM**, 2010.

[2] Anusha Ravula and Byrav Ramamurthy, "*A Tabu Search Approach for Joint Scheduling of Resources in a Lambda Grid Network*," IEEE **GLOBECOM**, 2010.

[3] Pragatheeswaran Angu and Byrav Ramamurthy, "*Experiences with dynamic circuit creation in a regional network testbed*," IEEE **INFOCOM HSN** Workshop, 2011.

[4] Pragatheeswaran Angu, Mukesh Subedee, Byrav Ramamurthy, Xi Yang & Tom Lehman, "*Inter-Domain Dynamic Circuit Network Creation*", In IEEE **GLOBECOM**, December 2010, Miami, FL. (*demo*).

[5] Lin Liu and Byrav Ramamurthy, "*Rightsizing Bundle Link Capacities for Energy Savings in the Core Network,*" IEEE **GLOBECOM**, 2011.

[6] Lin Liu and Byrav Ramamurthy, *"A Dynamic Local Method for Bandwidth Adaptation in Bundle Links to Conserve Energy in Core Networks*," in IEEE **ANTS**, 2011.

[7] Byrav Ramamurthy, K.K. Ramakrishnan, R.K. Sinha, *"Cost and Reliability Considerations in Designing the Next-Generation IP over WDM Backbone Networks,"* **ICCCN** 2011.

Acknowledgments: Joint work at UNL with Netgroup students: P. Angu, M. Subedee, A. Ravula, L. Liu, J. Mukherjee, A. Kolasani, A. Lara.